Les 5. Analyse van 2 variabelen

Onderzoekstechnieken

Jens Buysse **Wim De Bruyn** **Bert Van Vreckem**

AJ 2018-2019

HO GENT

What’s on the menu today?

Bivariate Analyse
    Bivariate Analyse: grafisch

Kruistabellen en Cramér’s V

Grafieken voor kruistabellen

Lineaire Regressie

Correlatiecoëfficiënt en determinatiecoëffciënt

HO GENT

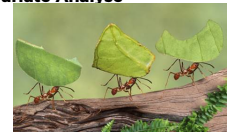
Bivariate Analyse

HO
GENT

Bivariate Analyse

HO GENT

Bivariate Analyse



Deze slide wordt gebruikt om een voorbeeld te geven van een minder triviaal verband tussen variabelen: Ant Colony optimization. Verbanden tussen variabelen zouden dus kunnen zijn:

- Aantal obstakels tussen nest en voedeselbron
- Algoritme gebruikt om feromonen weg te nemen / te plaatsen
- Vorm van de obstakels tussen nest en voedselbron
- ...

Voorbeeld

Tevredenheidsonderzoek campusrestaurant

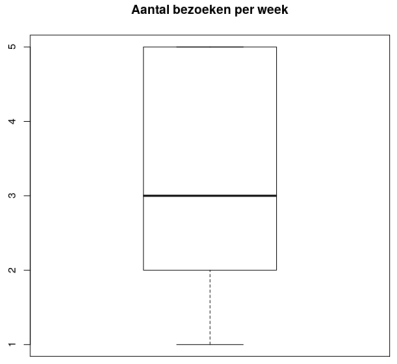
- Hoe vaakt bezoekt men het restaurant?
- Is er een verschil in uitgaven tussen student en medewerker?
- Is er een verband tussen het aantal dagen dat men bezoekt en bedrag dat men wekelijks besteedt?

R Code: zie `cursus/data/catering_hogeschool.R`

HO GENT

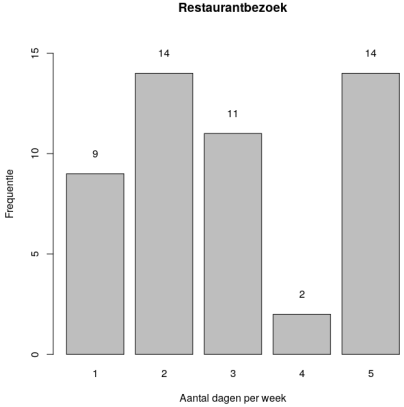
Hoe vaakt bezoekt men het restaurant?

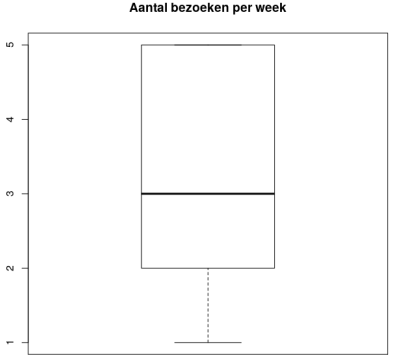
Statistiek	Waarde
Mean	2.96
Median	3
Mode	2
Stdev	1.484
Variantie	2.202
Range	4
Q_1	2
Q_2	3
Q_3	5



HO GENT

Hoe vaakt bezoekt men het restaurant?

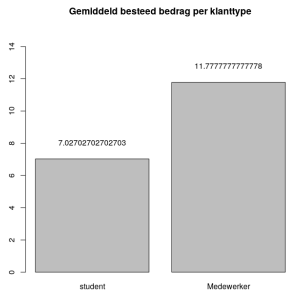




HO GENT

Student vs werknemer

- Enkelvoudig staafdiagram (van gemiddelde per categorie)
- Boxplot

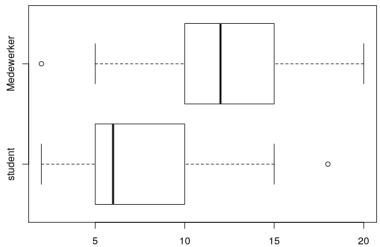


Let op! Onvoldoende om significant verschil aan te tonen!

HO GENT

Student vs werknemer

- Enkelvoudig staafdiagram (van gemiddelde per categorie)
- Boxplot



HO GENT

Afhankelijke en onafhankelijke variabele

HO GENT

Afhankelijke en onafhankelijke variabele



Onderzoeken die hier gevoerd zijn:

- Invloed van alcoholinname op leervermogen van vleermuizen (Drinking and Flying: Does Alcohol Consumption Affect the Flight and Echolocation Performance of Phyllostomid Bats?)
- Arnd Leike of the Ludwig Maximilians University receives one of the Ig Nobel awards - which are given for research that cannot or should not be repeated - for demonstrating that beer froth obeys the mathematical law of exponential decay.

Onderzoek academiejaar 2013-2014

HO GENT

Onderzoek academiejaar 2013-2014



Studenten moesten onderzoeken of er een verband was tussen vallen van boterham op boterzijde en hoogte e.a., of verband was tussen het aantal onpare sokken en andere fenomenen zoals je eigen was doen, veel sporten al dan niet ...

Kruistabellen en Cramér's V

HO
GENT

Kruistabellen

Is er een verschil in waardering in het assortiment tussen mannen en vrouwen?

	Vrouw	Man	Totaal
Goed	9	8	17
Voldoende	8	10	18
Onvoldoende	5	5	10
Slecht	0	4	4
Totaal	22	27	49

HO GENT

Kruistabellen: percenteren

Is er een verschil in waardering in het assortiment tussen mannen en vrouwen?

	Vrouw	Man	Totaal	Vrouw %	Man%	Totaal
Goed	9	8	17	41%	30%	35%
Voldoende	8	10	18	36%	37%	37%
Onvoldoende	5	5	10	23%	18%	20%
Slecht	0	4	4	0%	15%	8%
Totaal	22	27	49	100%	100%	100%

HO GENT

Kruistabellen: verschil bepalen

Is er een verschil in waardering in het assortiment tussen mannen en vrouwen?

	Vrouw	Man	Totaal	Vrouw %	Man%	Totaal
Goed	$9 - 7.63$	$8 - 9.36$	17	41%	30%	35%
Voldoende	$8 - 8.08$	$10 - 9.91$	18	36%	37%	37%
Onvoldoende	$5 - 4.48$	$5 - 5.51$	10	23%	18%	20%
Slecht	$0 - 1.79$	$4 - 2.20$	4	0%	15%	8%
Totaal	22	27	49	100%	100%	100%

HO GENT

Kruistabellen: kwadrateren en normeren

Is er een verschil in waardering in het assortiment tussen mannen en vrouwen?

	Vrouw	Man	Totaal	Vrouw %	Man%	Totaal
Goed	0.2	0.2	17	41%	30%	35%
Voldoende	0	0	18	36%	37%	37%
Onvoldoende	0.1	0	10	23%	18%	20%
Slecht	1.8	1.5	4	0%	15%	8%
Totaal	22	27	49	100%	100%	100%

$$\chi^2 = 3.811, V = 0.279$$

HO GENT

Cramér's V

Cramér's V is een maat die aanduidt hoe sterk de samenhang is tussen twee kwalitatieve variabelen. Dit getal ligt altijd tussen 0 en 1

Waarde	Interpretatie
0	geen samenhang
0.1	zwakke samenhang
0.25	redelijk sterke samenhang
0.5	sterke samenhang
0.75	zeer sterke samenhang
1	volledige samenhang

HO GENT

Voorbeeld 2: voorkeur automerk en geslacht

	Mercedes	BMW	Porsche	Alfa Romeo	Totaal
Mannen	10	10	20	20	60
Vrouwen	20	5	15	0	40
Totaal	30	15	35	20	100

Het lijkt alsof de automerken niet gelijkelijk gewaardeerd worden door mannen en vrouwen.

$$\chi^2 = 22.619$$

$$V = \sqrt{\frac{22.169}{100.(2-1)}} = 0.476$$

HO GENT

Voorbeeld2: voorkeur automerk en geslacht

geslacht * automerk Crosstabulation

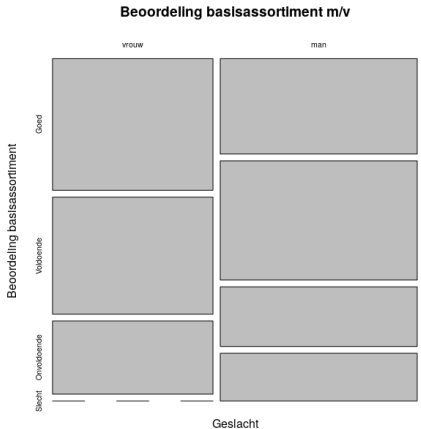
			automerk				
			Mercedes	BMW	Porshe	Alfa Romeo	Total
geslacht	man	Count	10	10	20	20	60
		Expected Count	18.0	9.0	21.0	12.0	60.0
		% within geslacht	16.7%	16.7%	33.3%	33.3%	100.0%
		% within automerk	33.3%	66.7%	57.1%	100.0%	60.0%
		Std. Residual	-1.9	.3	-.2	2.3	
	vrouw	Count	20	5	15	0	40
		Expected Count	12.0	6.0	14.0	8.0	40.0
		% within geslacht	50.0%	12.5%	37.5%	0.0%	100.0%
		% within automerk	66.7%	33.3%	42.9%	0.0%	40.0%
		Std. Residual	2.3	-.4	.3	-2.8	

HO GENT

Grafieken voor kruistabellen

HO
GENT

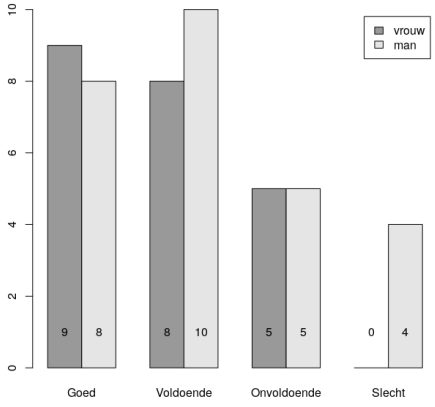
Visuele voorstelling van kruistabelen



HO GENT

Visuele voorstelling van kruistabelen

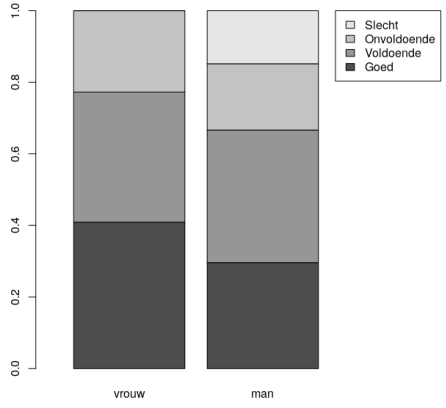
Geclusterde staafgrafiek



HO GENT

Visuele voorstelling van kruistabelen

Rependiagram



HO GENT

Lineaire Regressie

HO
GENT

Lineaire regressie

Bij regressie gaan we proberen een consistente en systematische koppeling tussen de variabelen te vinden.

1. **Monotoon:** algemene richting van de samenhang tussen de twee variabelen kan aangeduid worden (stijgend/dalend).
2. **Niet-monotoon:** aanwezigheid (of afwezigheid) van de ene variabele systematisch gerelateerd aan de aanwezigheid (of afwezigheid) van een andere variabele.

HO GENT

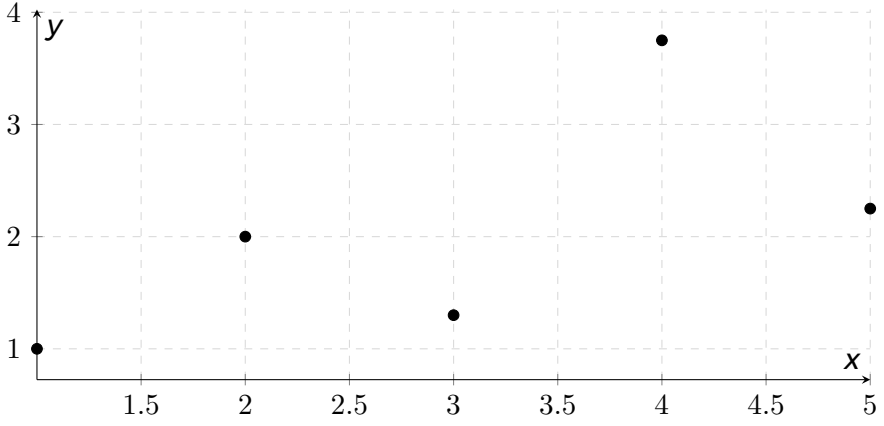
Lineaire regressie

Lineair verband: een rechtlijnige samenhang tussen een onafhankelijke en afhankelijke variabele, waarbij kennis van de onafhankelijke variabele kennis over de afhankelijke variabele geeft.

- Aanwezigheid
- Richting: dalend of stijgend?
- Sterke van het verband: sterk, gematigd, niet bestaand ...

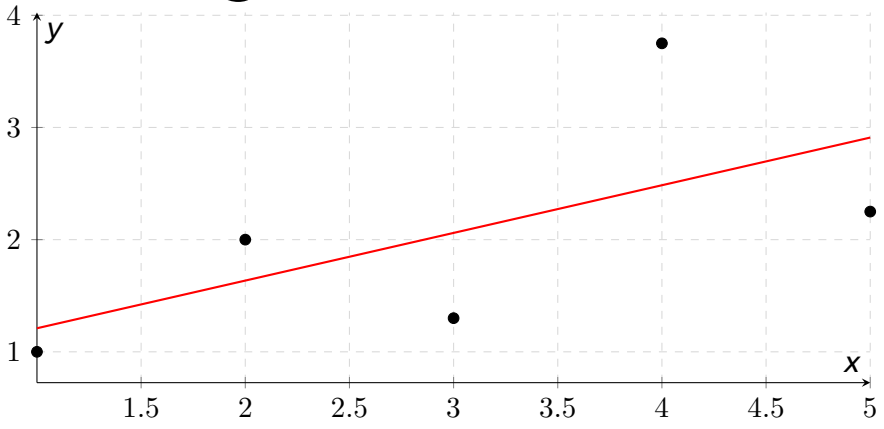
HO GENT

Lineaire regressie



HO GENT

Lineaire regressie



HO GENT

Kleinste kwadratenmethode: voorbeeld

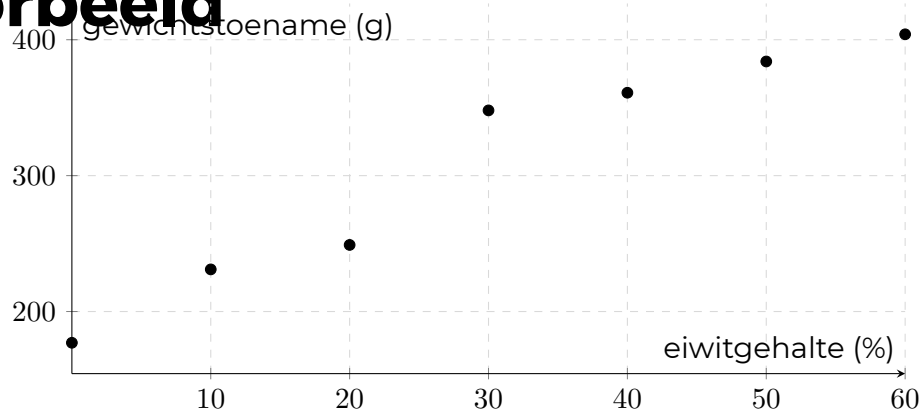
HO
GENT

Kleinste kwadratenmethode: voorbeeld

Eiwitgehalte%	Gewichtstoename (gram)
0	177
10	231
20	249
30	348
40	361
50	384
60	404

HO GENT

Kleinste kwadratenmethode: voorbeeld



HO GENT

Kleinste kwadratenmethode: voorbeeld

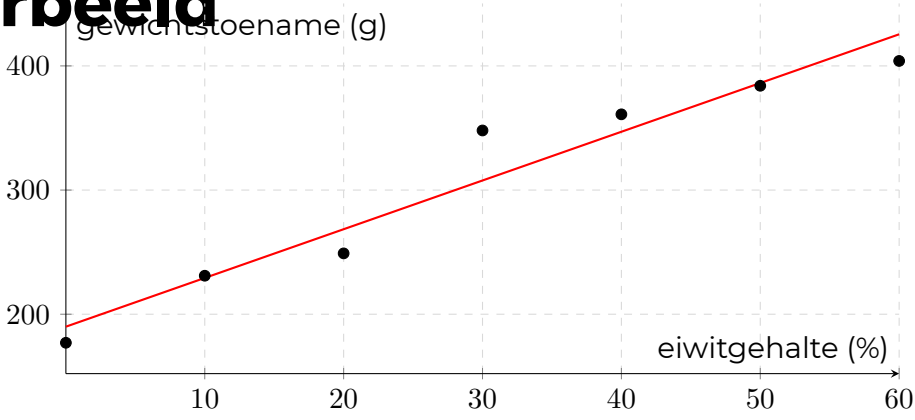
x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
0	177	-30	-130,71	3921,3	900
10	231	-20	-76,71	1534,2	400
20	249	-10	-58,71	587,1	100
30	348	0	40,29	0	0
40	361	10	53,29	532,9	100
50	384	20	76,29	1525,8	400
60	404	30	96,29	2888,7	900
				10990	2800

Tabel: Berekeningen die nodig zijn voor het toepassen van de kleinste kwadratenmethode.

HO GENT

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x-\bar{x})^2} = \frac{10990}{2800} = 3.925$$

Kleinste kwadratenmethode: voorbeeld



Correlatiecoëfficiënt en determinatiecoëffciënt

HO GENT

Pearson correlatiecoëfficiënt en determinatiecoëfficiënt

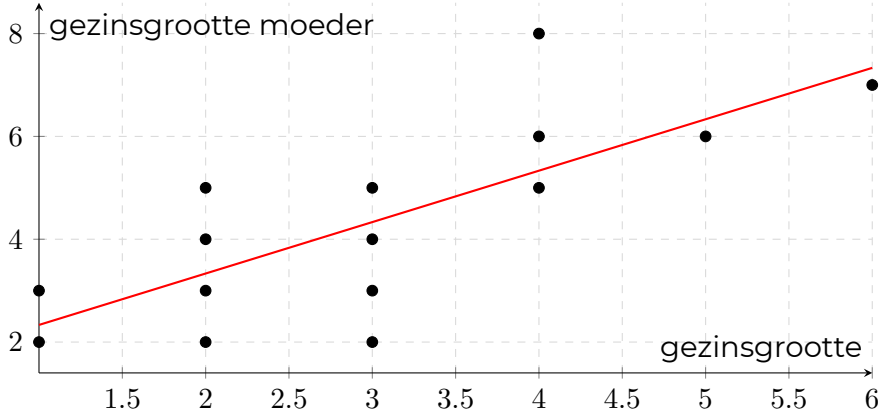
De **Pearson correlatiecoëfficiënt** is een maat voor de sterkte van de lineaire samenhang tussen x en y

De **determinatiecoëfficiënt** verklaart het percentage van de variantie van de waargenomen waarden t.o.v. de regressierechte.

HO GENT

Covariantie

We plotten de gezinsgrootte van 15 families tot de gezinsgrootte van de moeder toen ze klein was.

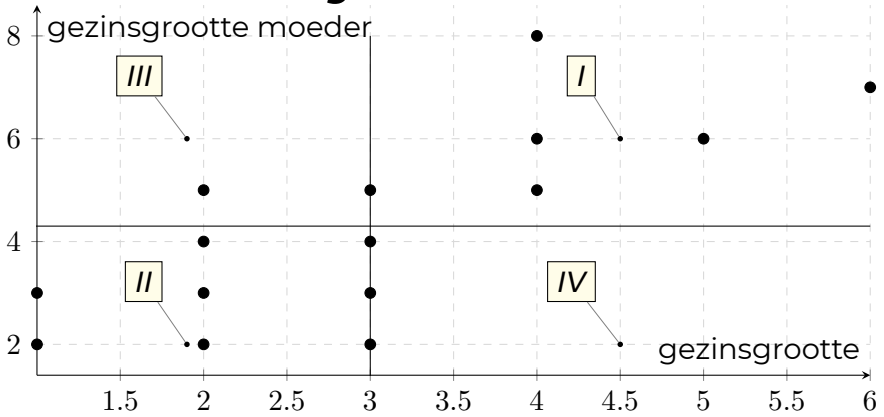


We hebben

$\overline{x} = 3$ en $\overline{y} = 4.3$.

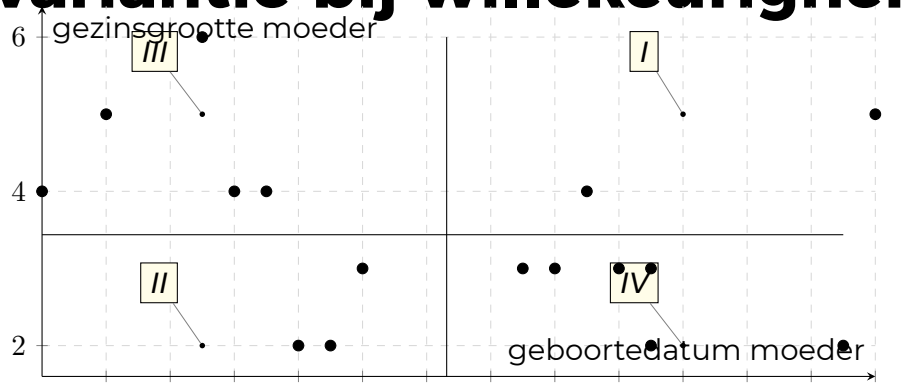
HO GENT

Covariantie bij lineair verband



HO GENT

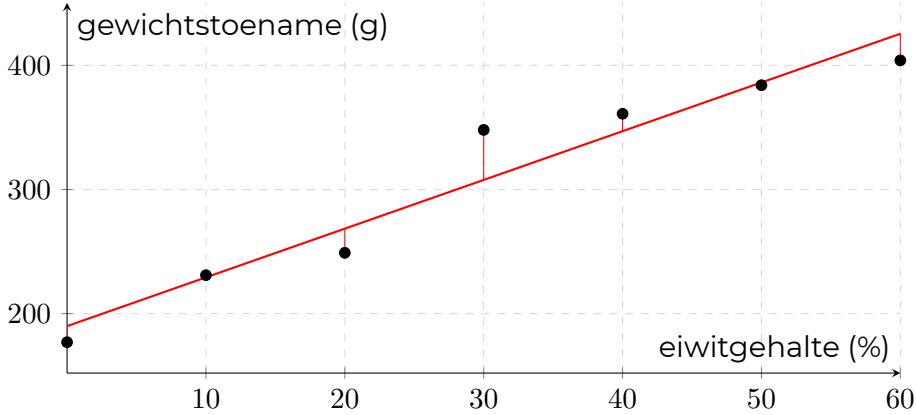
Covariantie bij willekeurigheid



We vinden $\overline{x} = 1942.625$ en $\overline{y} = 3.4375$.

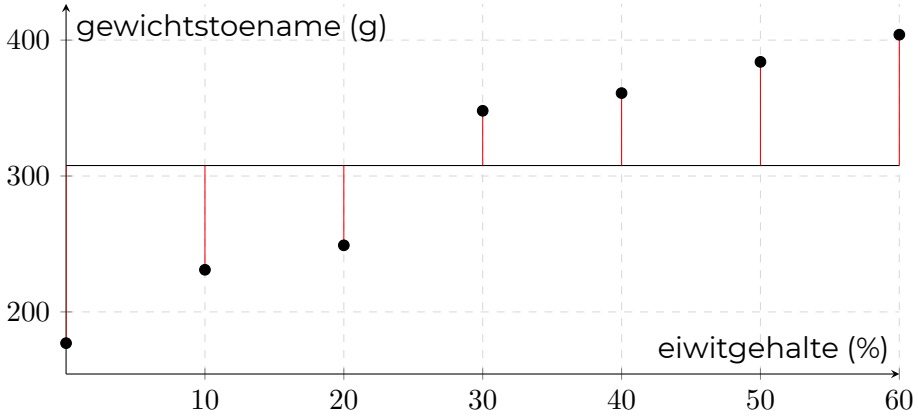
HO GENT

Determinatiecoëfficiënt



Figuur: Deviaties tot de regressierechte: aanname ***x*** geeft extra informatie voor het voorspellen van ***y***.

HO GENT



Figuur: Deviaties tot de gemiddelde van y: aanname ***x*** geeft geen informatie voor het voorspellen van ***y*** ($\bar{y} = 307.71$).

HO GENT

Correlatiecoëfficiënt en determinatiecoëffciënt

R	R^2	Verklaarde variantie	Interpretatie
$< 0,3$	$< 0,1$	$< 10\%$	zeer zwak
$0,3 - 0,5$	$0,1 - 0,25$	$10 - 25\%$	zwak
$0,5 - 0,7$	$0,25 - 0,5$	$25 - 50\%$	matig
$0,7 - 0,85$	$0,5 - 0,75$	$50 - 75\%$	sterk
$0,85 - 0,95$	$0,75 - 0,9$	$75 - 90\%$	zeer sterk
$> 0,95$	$> 0,9$	$> 90\%$	uitzonderlijk(!)

HO GENT

Sterkte verband rendieren

$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
-30	-130.714	3921.429
-20	-76.7143	1534.286
-10	-58.7143	587.1429
0	40.28571	0
10	53.28571	532.8571
20	76.28571	1525.714
30	96.28571	2888.571

$$\sum_{i}^{n}(x-\bar{x})(y-\bar{y}) = 10990$$

$$Cov = \frac{10990}{7} = 1570$$

$$\sigma_x = 20$$

$$\sigma_y = 81.03$$

$$R = \frac{1570}{20 \times 81.03} = 0.96$$

$$R^2 = 0.93$$

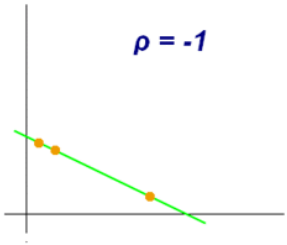
HO GENT

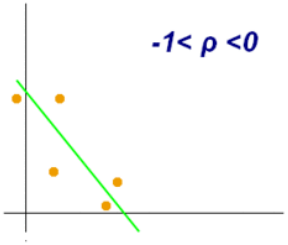
Overwegingen

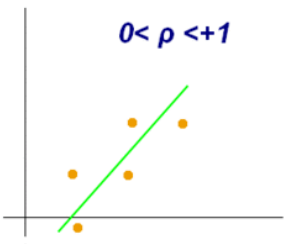
- Bij de correlatiecoëfficiënt wordt er alleen naar het verband tussen twee variabelen gekeken. Er wordt niet gekeken naar interacties met andere variabelen.
- Er wordt bij de correlatiecoëfficiënt expliciet niet uitgegaan van een oorzaak-en gevolg verband
- De product-momentcorrelatiecoëfficiënt van Pearson drukt slechts lineaire verbanden uit

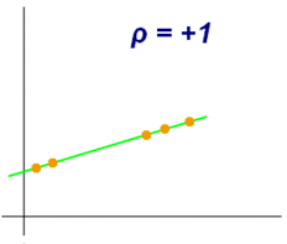
HO GENT

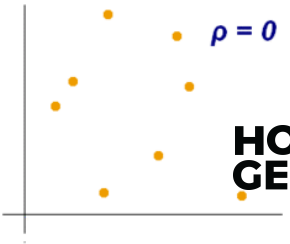
Verband regressierechte en correlatecoëfficiënt











HO GENT